Assignment 2

2.1

a) $$\underset{w}{\text{minimize}}\ f(w) = \frac{1}{N}\sum_{i\in[N]} f_i(w) + \lambda\|w\|_2^2,$$

Where $f_i(w) = \log\left(1+\exp(-y_i w^T x_i)\right)$

a) f is not Lipschitz continuous, because the term $\|w\|_2^2$ is not Lipschitz continuous.

To further clarify, let

$$g(w) = \|w\|_2^2 = w^T w$$

Then; $\nabla g = 2w \Longrightarrow \|\nabla g\|$ is not bounded.

~~$\therefore g(w) = \|w\|_2^2$ is not Lipschitz continuous.~~

~~Thus f(w) is not Lips~~

Thus f is not Lipschitz continuous.

b) $\|\nabla f_i(w_2) - \nabla f_i(w_1)\|$

$$= \left\| \frac{-y_i x_i e^{-y_i w_2^T x_i}}{1+e^{-y_i w_2^T x_i}} + \frac{y_i x_i e^{-y_i w_1^T x_i}}{1+e^{-y_i w_1^T x_i}} \right\|$$

$$= |y_i|\,\|x_i\| \left\| \frac{e^{-y_i w_1^T x_i}}{1+e^{-y_i w_1^T x_i}} - \frac{e^{-y_i w_2^T x_i}}{1+e^{-y_i w_2^T x_i}} \right\|.$$

$$= |y_i|\,\|x_i\| \left\| \frac{1}{1+e^{y_i w_2^T x_i}} - \frac{1}{1+e^{y_i w_1^T x_i}} \right\|$$

let $h(w) = \dfrac{1}{1+e^{y_i w^T x_i}}$ ~~(This is the~~

$h(w)$ is continuous and differentiable.

$\therefore$ By mean value theorem, $\cancel{\exists z \in \mathbb{R}^n \text{ s.t}}$ $\exists z \in (w_1, w_2)$ s.t

$$\cancel{\nabla f(z)}$$

$$h(w_2) - h(w_1) = \nabla h(z)^T (w_2 - w_1)$$

$$\Longrightarrow \| h(w_2) - h(w_1) \| \leq \| \nabla h(z) \| \, \| w_2 - w_1 \| \quad \text{—— } (2)$$

We have $\nabla h(w) = \dfrac{-y_i x_i e^{y_i w^T x_i}}{(1+e^{y_i w^T x_i})^2}$

$\therefore$ (2) $\Longrightarrow$

$$\| h(w_2) - h(w_1) \| \leq \| y_i x_i \| \left\| \frac{e^{y_i z^T x_i}}{(1+e^{y_i z^T x_i})^2} \right\| \| w_2 - w_1 \|$$

$$= |y_i| \, \|x_i\| \left\| \frac{1}{(1+e^{-y_i z^T x_i})(1+e^{y_i z^T x_i})} \right\| \times \| w_2 - w_1 \|$$

$$\leq |y_i| \, \|x_i\| \, \| w_2 - w_1 \|$$

$\therefore$ (1) $\Longrightarrow$

$$\| \nabla f_i(w_2) - \nabla f_i(w_1) \| = |y_i| \, \|x_i\| \, \| h(w_2) - h(w_1) \|$$

$$\leq |y_i| \, \|x_i\| \, |y_i| \, \|x_i\| \, \| w_2 - w_1 \|$$

$$= \underbrace{|y_i|^2 \|x_i\|^2}_{L} \| w_2 - w_1 \|$$

~~$\therefore$ small~~

$\therefore$ a small L for f_i is $|y_i|^2 \|X_i\|^2$.

* ∇f is also Lipschitz continuous.

proof:

$$\|\nabla f(w_2) - \nabla f(w_1)\|$$

$$= \Big\| \frac{1}{N} \sum_{i \in [N]} \nabla f_i(w_2) + \cancel{2\|w_2\|^2}\ \lambda \cdot 2 w_2$$

$$- \left[\frac{1}{N} \sum_{i \in [N]} \nabla f_i(w_1) + \lambda \cdot 2 w_1\right] \Big\|$$

$$= \Big\| \frac{1}{N} \sum_{i \in [N]} \left(\nabla f_i(w_2) - \nabla f_i(w_1)\right) + 2\lambda (w_2 - w_1) \Big\|$$

$$\leq \frac{1}{N} \sum_{i \in [N]} \|\nabla f_i(w_2) - \nabla f_i(w_1)\| + 2\lambda \|w_2 - w_1\|$$

$$\leq \frac{1}{N} \left(\sum_{i \in [N]} |y_i|^2 \|X_i\|^2\right) \|w_2 - w_1\| + 2\lambda \|w_2 - w_1\|$$

$$= \underbrace{\left[\frac{1}{N} \left(\sum_{i \in [N]} |y_i|^2 \|X_i\|^2\right) + 2\lambda\right]}_{L_1} \|w_2 - w_1\|$$

$\implies$ f is L_1-smooth

c) f is strongly convex, because $\lambda \|W\|^2$ is strongly convex and $\frac{1}{N} \sum_{i \in [N]} f_i(w)$ is convex

Let $g(w) = \lambda \|w\|^2$

$= \lambda w^T w$

$\implies \nabla g(w) = 2 \lambda w$

$\implies \nabla^2 g(w) = 2 \lambda I$

$\implies g$ is strongly convex with $\mu = 2\lambda$.

$\therefore$ f is also 2λ strongly convex.

2-) We want to find a relationship between (1a), (1b), (2) and given relation and prove that it holds

If we open the variance relation,

$$Var_{\xi_t}[g(\underline{w}_t;\xi_t)] = E\{[g(\underline{w}_t;\xi_t) - E\{g(\underline{w}_t;\xi_t)\}]^2\} =$$

$$= E\{\|g(\underline{w}_t;\xi_t)\|_2^2 - E\{g(\underline{w}_t;\xi_t)^T\} g(\underline{w}_t;\xi_t) - g(\underline{w}_t;\xi_t)^T E\{g(\underline{w}_t;\xi_t)\}$$

$$+ \|E\{g(\underline{w}_t;\xi_t)\}\|_2^2\}$$

$$\leq M + M_v \|\nabla f(\underline{w}_t)\|_2^2$$

Therefore,

$$E\{\|g(\underline{w}_t;\xi_t)\|_2^2\} \leq 2E\{E\{g(\underline{w}_t;\xi_t)\}^T g(\underline{w}_t;\xi_t)\}$$

$$- \|E\{g(\underline{w}_t;\xi_t)\}\|_2^2 + M + M_v\|\nabla f(\underline{w}_t)\|_2^2$$

$$= \|E\{g(\underline{w}_t;\xi_t)\}\|_2^2 + M + M_v \|\nabla f(\underline{w}_t)\|_2^2$$

Using (16),

$$\| E\{ g(w_k ; \xi_k) \} \|_2 \leq c_0 \| \nabla f(w_k) \|_2$$

and taking square of the relation, we obtain

$$E\{ \| g(w_k ; \xi_k) \|_2^2 \} \leq M + (c_0^2 + M_V) \cdot \| \nabla f(w_k) \|_2^2$$

Therefore, we have $\alpha = M$, $\beta = c_0^2 + M_V$ and the relation is satisfied.

3-) We want to prove the given relation for SGD with non-convex objective function and square summable but not summable step-size.

As step size is square summable, $\{\alpha_k\} \to 0$.

There fore, we may take $\alpha_k L \cdot M_G \leq \mu$. $\forall k \in \mathbb{N}$

We use the approach in references and look at,

$$E\{f(w_{k+1})\} - E\{f(w_k)\} \leq -(\mu - \frac{1}{2}\alpha_k L \cdot M_G)\, \alpha_k E\{\|\nabla f(w_k)\|_2^2\} + \frac{1}{2}\alpha_k^2 L \cdot M.$$

$$\leq -\frac{1}{2}\mu \alpha_k E\{\|\nabla f(w_k)\|_2^2\} + \frac{1}{2}\alpha_k^2 L \cdot M.$$

As in p 3.22. of lecture notes, we sum the relations, and obtain,

$$f_{inf} - E\{f(w_1)\} \leq E\{f(w_{K+1})\} - E\{f(w_1)\}$$
$$\leq -\frac{1}{2}\mu \sum_{k=1}^{K} \alpha_k E\{\|\nabla f(w_k)\|_2^2\} + \frac{1}{2} L \cdot M \cdot \sum_{k=1}^{K} \alpha_k^2$$

Further using algebraic manipulations,

$$\sum_{k=1}^{K} \alpha_k E\{\|\nabla f(w_k)\|_2^2\} \leq \frac{2(E\{f(w_1)\} - f_{inf})}{\mu} + \frac{LM}{\mu} \sum_{k=1}^{K} \alpha_k^2$$

as α_k are square summable, we have

$$\sum_{k=1}^{K} \alpha_k^2 < \infty$$

It converges to a finite limit as K increases:

Which shows that the first relation holds.

As α_k is not summable,

$$\sum_{k=1}^{k} \alpha_k = \infty$$

Tends to infinity. Therefore,

$$E\left\{\left(\frac{1}{\sum_{k=1}^{K} \alpha_k}\right) \cdot \sum_{k=1}^{K} \alpha_k \|\nabla f(w_k)\|_2^2\right\} \xrightarrow{K \to \infty} 0$$

As the summation is in denominator.